doi:10.16055/j.issn.1672-058X.2015.0002.009

一种基于数据挖掘的零售业客户细分方法研究*

蔡玖琳, 张　磊**, 张秋三
(青岛大学 管理科学与工程学院,山东 青岛 266001)

摘　要:针对零售业客户细分指标粗糙和方法精准性低的问题,提出一种基于数据挖掘聚类分析的零售业客户细分方法;方法构建了一套基于 RFM 的多指标客户细分指标体系,采用熵值法赋予指标权重,进而使用 *K*-Means 算法进行客户细分;实证研究结果表明:方法在客户行为特征区分能力和聚类紧凑性方面均优于传统基于 RFM 的细分方法,方法可行、有效,能够更好地解决零售业客户细分问题,提升客户关系管理和营销决策质量.

关键词:客户细分;RFM;熵值法;*K*-Means
中图分类号:TP392　　**文献标识码**:A　　**文章编号**:1672-058X(2015)02-0043-06

作为国民经济的重要组成部分,零售业发展迅猛.但是随着竞争的加剧和开放的深入,零售业的利润日渐微薄,迫切需要寻找新的利润增长点.对于现在供过于求的零售业来说,如何赢得和保留客户并将客户价值最大化变得尤为重要,也日渐成为企业最为关注的问题之一.零售业存在大量的客户数据和销售数据,这些数据的数量随着时间的推移呈现爆炸式增长.信息技术的飞速发展和大数据时代的到来使企业能够借助数据挖掘等技术,充分利用这些海量数据对客户进行细分,帮助企业制定决策,在更好地满足客户需求的同时,为企业获取持续忠诚的客户和更高的利润.

企业自身资源的局限性,一定程度上决定了企业不可能达到让所有客户都满意的目标.依据 80/20 法则,企业最具赢利性的 20%的客户创造了企业 80%的利润[1].通过对会员客户进行细分,制定更加贴近客户需求的营销策略,能够使企业更有效地利用自身资源,提升客户满意度和客户收益.相对其他细分方法,数据挖掘方法更加准确和科学,在传统基于 RFM(Recency, Frequency, Monetary)的方法基础上,提出一种基于聚类算法的多指标客户细分方法,并结合具体零售企业实例验证其有效性.

1　客户细分方法研究现状

客户细分方法主要有基于经验的方法、统计分析方法以及非统计分析方法 3 种.前两种方法只能进行相对简单的客户细分,已经无法满足企业的需求.随着 20 世纪 90 年代数据挖掘技术的兴起,复杂的零售业客户细分有了全新的非统计分析方法.数据挖掘技术中应用于客户细分最为广泛的是聚类技术,使用聚类分析来对客户进行分组,使组内成员具有类似的特征,而组间成员之间的差异较大,以此加强客户关系管理的商务策略,帮助改进服务质量,提高顾客满意度和忠诚度[2].叶孝明、黄祖庆[3](2006)、周颖[4](2007)、Cheng[5](2009)、徐翔斌[6](2012)等在研究中都采用了聚类技术来进行客户细分,衡量客户价值,效果显著.其中,最

收稿日期:2014-06-30;**修回日期**:2014-09-10.
* **基金项目**:国家自然科学基金资助(71273148).
作者简介:蔡玖琳(1990-),女,湖北黄冈人,硕士研究生,从事数据挖掘与数据分析研究.
** **通讯作者**:张磊(1978-),男,山东即墨人,副教授,博士,从事数据挖掘研究.E-mail:qduzhanglei@ qdu.edu.cn.

常用到的算法是基于划分的 K-Means 聚类算法.

现有客户细分方法中所用指标体系的构建经历了传统客户细分和基于客户价值的细分两大阶段.传统客户细分更多基于地理位置和社会经济学等特征进行细分.Lazer(1964)提出以生活方式为背景识别和细分客户,Mitchell(1983)提出了一种基于社会阶层、生活方式和个人特征的可概括的心理细分模型.随着社会发展和经济全球化的推进,这些因素不足以区分客户行为,细分效果也大打折扣.于是客户细分指标的构建进入基于客户价值的细分阶段,最具代表性的就是 Hughes 提出的 RFM 分析方法[7].其中,R(recency,近度),指的是客户最后一次交易行为与当前的时间间隔,间隔时间越短,R 赋值越大;F(frequency,频度),指的是某一特定的时期内客户的交易次数;M(monetary,额度),指的是某一特定时期内客户的交易金额.Chang 等人指出,R 越大,F 越大的客户与企业达成新交易的可能性也越大,M 越大的客户再次响应企业的产品和服务的可能性越大[8].传统的 RFM 分析方法依据以上 3 个行为变量对每个客户打分,然后计算 3 个指标的乘积,根据最后的结果对所有的客户进行排序,再按照一定的比例进行分类,最后针对不同的客户群体制定不同的服务策略.RFM 所涉及的变量都与客户的行为相关,容易从相关数据库中提取,并且能够预测客户的购买行为[9].但是 RFM 分析过程复杂,耗时长,细分结果可能会产生过多的客户群体,如果给每一种变量赋予 5 个值,就会得到 125 个细分客户群.

不同的客户细分指标对于聚类结果的重要性可能不同,因此需要视情况对指标赋予不同的权重.国内外学者对于细分指标的权重问题进行了诸多研究,其中主要分为两大类:以层次分析法、特征值法为代表的主观评价赋值法和以极差法、熵值法为代表的客观评价赋值法.两类赋权方法各有千秋,主观评价以人的主观判断为依据,受到参与评价者自身判别能力和倾向的影响,可能在评价结果中产生一定的随意性;而客观评价更加注重数学理论知识的应用,不受决策者主观意识的影响,对于结果的产生也可能存在偏差[10].

2 基于聚类分析的零售业客户细分方法

2.1 客户细分指标体系构建

传统的 RFM 分析方法在每一个维度上只有一个指标,不能全面衡量客户的行为特征,结合未来客户细分所呈现的多维度趋势,在传统 RFM 分析的基础上,对 RFM 进行多指标化处理,构建刻画客户消费行为特征的指标体系,如表 1 所示.

表 1 客户细分指标体系

指标体系	细分类别
消费间隔时间(R)	R_1:客户近度与全部客户平均近度之比
	R_2:某客户最新一年近度与自身历史近度的比值
消费频次(F)	F_1:某客户消费次数与全部客户平均消费次数的比值
	F_2:某客户最近一年消费次数与自身历史消费次数的比值
消费金额(M)	M_1:某客户消费金额与全部客户平均消费金额的比值
	M_2:某客户最新一年消费金额与自身历史消费金额的比值
已消费物品种类(C)	C:客户不重复消费类别

客户细分指标体系将 RFM 每个维度扩充为两个指标,分别从宏观和微观两个角度来观察客户行为,另外增加了对已消费物品种类的考虑.表 1 中,R_1,F_1 和 M_1 为从宏观角度刻画客户的长期行为的指标,R_2,F_2 和 M_2 为从微观角度来刻画客户的近期动态的指标,C 为“已消费物品种类”指标.从宏观角度来讲,将客户自身指标与全体客户对应指标的平均值比较,可以更清楚地确定客户在整体中的相对位置;从微观角度来讲,将客户的最近一年指标与历史指标对照,可以从自身的角度观察客户的近期行为取向;而已消费物品种类代表了对客户进行交叉销售的可能性,从一定程度上反映了客户的潜在价值.同时考虑以上 3 个方面,该客户细分指标体系可以更全面地观察客户的行为特征.

2.2　客户细分指标赋权

从细分目的和所构建的细分指标体系角度考虑,为了消除主观差异,采用客观的赋权方法——熵值法为客户细分指标体系的各个指标赋予权重,具体步骤如下:

1) 原始数据标准化.假设存在 m 个客户,n 个细分指标,那么原始数据集为

$$\boldsymbol{X}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

对于其中值越大越好的正向指标而言,其标准化公式为

$$x'_{ij}=\frac{x_{ij}-\min\{x_j\}}{\max\{x_j\}-\min\{x_j\}} \tag{1}$$

对于其中值越小越好的负向指标而言,其标准化公式如下

$$x'_{ij}=\frac{\min\{x_j\}-x_{ij}}{\max\{x_j\}-\min\{x_j\}} \tag{2}$$

2) 计算第 i 个客户第 j 个指标的特征比重:

$$Y_{ij}=\frac{x'_{ij}}{\sum_{i=1}^{m}x'_{ij}} \tag{3}$$

3) 计算指标 j 的信息熵:

$$e_j=-k\sum_{i=1}^{m}(Y_{ij}\mathrm{Ln}Y_{ij}) \tag{4}$$

其中 $k>0$,设定 $k=1/\mathrm{Ln}m$,m 为包含的客户人数,即整个数据矩阵的行数.

4) 计算指标的信息效用度.对于给定的指标 j,x'_{ij}的差异越小,则 e_j 的值越大,指标的作用也越小,x'_{ij}的差异越大,则 e_j 的值越小,指标的作用也越大.如果 x'_{ij}的值相等,则 e_j 达到最大值 1,这种情况下,这个指标的存在就没有意义,应该从指标体系中剔除[10].因此,定义指标的信息效用度为

$$d_j=1-e_j \tag{5}$$

信息效用度的值越大,应该越重视这个指标在整个指标体系当中的作用.

5) 确定各指标的权重,重新计算各指标赋权之后的值各指标的权重构成的权重向量为

$$\boldsymbol{w}=\{w_1,\ w_2,\ \cdots,\ w_n\}\ ,0\leqslant w_j\leqslant 1,j=1,2,\cdots,n,\ w_j=\frac{d_j}{\sum_{j=1}^{n}d_j} \tag{6}$$

令 $\boldsymbol{W}=\begin{pmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{pmatrix}$,将权重赋予各自对应的指标,形成新的数据集,即

$$\boldsymbol{X'}=\boldsymbol{XW} \tag{7}$$

2.3　客户细分聚类算法

K-Means 算法是在实践中常用的聚类算法之一,广泛应用于多个行业中,大量研究表明该算法在处理海量数据上可以取得不错的效果.算法的核心思想,是通过迭代“质心”并根据样本与质心的距离把各个样本指派到各个簇当中去.将该算法应用到前面所得到的加权后的新数据集 $\boldsymbol{X'}$,主要步骤如下:

1) 确定初始质心.首先,由用户指定聚类结果中簇的个数 K,然后选取第一个样本,作为第一个质心;然后,计算其余所有样本到质心的欧氏距离,选择距离最大的那个样本作为第二个质心,重复以上步骤直至选出 K 个质心.

2) 将样本指派到不同的簇.在算法迭代过程中,样本被指派到离自己最近的质心所代表的簇,距离用欧氏距离的平方来表示,即样本 i 到质心 j 的距离为

$$d_{ij} = \|\boldsymbol{X}_i - \boldsymbol{C}_j\|^2 = \sum_{p=1}^{n}(x_{ip} - c_{jp})^2 \tag{8}$$

其中,$\boldsymbol{X}_i$ 是样本 i 所有指标形成的向量,$\boldsymbol{C}_j$ 是簇 j 的质心对应这些指标的向量,n 是指标的个数.

3) 更新质心.在迭代过程中,重新计算每个簇的质心,所得到的第 k 个簇的质心向量为

$$\overrightarrow{\boldsymbol{X}_k} = (x_{k1}, x_{k2}, \cdots, x_{kn}) \tag{9}$$

其中,指派样本后,第 k 个簇中的样本数量为 m_k,向量的第 j 个分量 x_{kj} 为

$$x_{kj} = \frac{\sum_{m=1}^{m_k} x_{mj}(k)}{m_k}$$

$x_{mj}(k)$ 指的是簇 k 中样本 m 的第 j 个指标的值.

4) 停止准则.*K*-Means 算法一般采用最大迭代次数或者差异容忍度作为控制算法是否终止的条件,此处采用前者,即通过设定最大迭代次数阈值作为算法终止条件.

3 客户细分实证研究

为了评估以上所提出的客户细分方法,选用谢邦昌编著的《数据挖掘基础与应用》一书所附某连锁超市VIP会员客户数据集进行测试[11].数据集包含32 810条会员信息,379 824条消费记录.对存在多个属性值缺失的会员信息予以删除,并对数据进行清洗等预处理,最终保留了32 699条会员信息记录以及378 744条消费记录.

1) 指标值计算.通过计算,得到所构建的客户细分指标体系中各指标值的描述统计如表2所示,其中 R_1,R_2 为负向指标,其余指标均为正向指标.

表2 指标特征描述统计

	最大值	最小值	均值	标准差
R_1	0.194 519	0.008 849	0.073 521	0.040 286
R_2	0.095 753	−0.036 06	0.013 55	0.019 367
F_1	7.097 623	0.027 725	0.321 133	0.272 097
F_2	0.243 759	−0.134 38	−0.001 3	0.013 946
M_1	0.326 765	0	0.013 561	0.012 576
M_2	0.437 398	−0.132 18	−0.000 79	0.011 365
C	13.994 18	0.241 279	2.276 562	1.500 04

2) 指标赋权.应用指标赋权方法计算各指标赋权之后的值,通过熵值法计算得到各个指标的权重为 $\boldsymbol{w}$ = (0.032 095 9, 0.005 318, 0.073 507, 0.028 624, 0.326 765, 0.003 548, 0.241 279) 将得到的权重赋予各自的属性,按式(7)计算得到新的数据集.

3) 对新数据集进行聚类,将客户分为4类,并根据客户的个人特征和消费特征对聚类结果进行分析,如表3所示.

表3 多指标客户细分结果

	人 数	会籍长度(众数)/年	平均已消费商品种类	平均近度	平均交易次数	平均消费金额/元	总金额/元	类内平均距离(欧氏)
C_1	4 658	2	18	355	23	20 343.97	94 762 214	0.009
C_2	1 431	4	27	307	42	39 944.27	57 160 248	0.022
C_3	8 518	3	11	434	13	11 720.12	99 831 942	0.012
C_4	18 092	3	5	490	6	5 547.08	1*E*+08	0.009

其中,会籍长度采用众数而非平均值来刻画,是因为会籍长度集中在单一的几个数值,差异性不大,采用众数来衡量更能体现聚类之间的差别.

第一类客户(C_2), VIP 客户,是企业的生力军.他们的总人数只占到整体客户的 4%,购买总额占了整个企业总额的 16%,并且消费种类繁多,购买频率高,购买金额非常大.但是,该类客户的年购买频率和年购买金额与其历史消费记录相比呈现一种缓慢下降的趋势.从客户与企业的交互特征来看,他们中大多数入会时间较长,企业在客户关系管理的过程中,应着重关注这部分能给企业带来巨大收益的客户,必要的时候可以为其进行个性化设计,进一步提升客户的忠诚度.例如将 VIP 客户作为交叉销售的重点对象,他们的平均消费种类在 27 种以上,对这类客户进行交叉销售可以获得相当不错的结果.

第二类客户(C_1),重要客户.这类客户人数占到整体的 14.2%,购买额度达到企业整体销售额度的 27%,平均消费种类高于整体水平,平均购买金额较高,频率较高.虽然购买力略逊于 VIP 客户,但是整体来说属于企业的忠诚客户,他们的会籍长度在 3 年及以上的达到了 66%,表明他们与企业之间建立了一种长久牢固的关系,企业在营销活动中应该注重保持与这类重要客户关系的维系,刺激客户消费,以求为企业带来更为长远稳定的收益.

第三类客户(C_3),一般重要客户.整体表现为平均消费金额和消费种类接近全体客户的平均水平,占总体 26%的客户购买金额占了企业销售总额的 28.4%,是销售收益贡献最高的群体.然而这个群体中,一半左右的客户的购买时间间隔较长,流失的可能性比较高.另一方面,这类客户中 68%的人入会时间不超过 3 年,与企业建立联系的时间相对较短,但是作为贡献率最高的客户群,企业应该重点发展此类客户,减少流失可能性.

第四类客户(C_4),一般客户.这类客户在总人数中占据了半壁江山,虽然最近一年的购买频率和购买金额相对其历史消费而言有走高的趋势,但是消费总额只占所有客户消费总额的 28%,其中 79%的客户入会时间不超过 3 年,整体表现为最后一次购买时间久远,购买金额小、频率低,会龄较短.企业可以适当降低在这类客户身上的资源投入,转移到其他客户群体身上来达到企业有限资源的有效利用.

4　与传统 RFM 的方法比较

采用前面同一个数据集,依据传统 RFM 细分指标对客户进行细分得到的结果如表 4 所示.从表 4 可以看出,只依据 RFM 分析方法对客户进行细分,得到的结果中 C_1,C_3 和 C_4 的各项特征中除了平均间隔时间差异较大,其余特征差异相对较小,总的来说可以将客户分为两大类:C_2 为一类,即企业的重要价值客户,C_1,C_3 和 C_4 为一类,即一般客户.并且聚类结果显示各个类内距离相对较大,而通过与表 3 对比发现,采用多指标客户细分方法得到的各个类之间差异较大,类内差异非常小,聚类效果更为紧凑.这表明所提出的方法在区分客户行为特征,进行客户细分方面明显优于传统基于 RFM 的细分方法,能够更好地为企业区分不同类型客户,制定差异化营销策略,使企业合理利用有限资源来提高客户忠诚度,提升企业自身价值.

表 4　传统 RFM 细分结果

	人　数	会籍长度(众数)/年	平均已消费商品种类	平均近度	平均交易次数	平均消费金额/元	总金额/元	类内平均距离(欧氏)
C_1	15 335	3	7	307	8	6 445.43	98 840 601	0.08
C_2	6 268	2	20	309	27	24 531.72	1.54*E*+08	0.13
C_3	6 796	3	8	564	10	9 435.69	62 124 973	0.1
C_4	4 300	5	6	973	7	8 228.31	35 381 721	0.11

5 结 语

提出了一种新的客户细分方法,从宏观和微观两个角度考虑,将传统指标扩充为4个类型7个指标,通过熵值法为指标赋权,并采用*K*-Means聚类算法进行客户细分.对某连锁超市会员客户进行细分的实证研究表明,方法在特征区分能力和聚类紧凑性方面比传统基于RFM的分析方法具有更佳的效果,可以帮助企业做出更准确高效的决策.但是,研究工作仍然存在一些不足的地方,例如*K*-Means聚类算法在初始聚类中心的选取上具有随机性、对离群点敏感等,这都将是下一步研究需要解决的问题.

参考文献:

[1] DUBOFF R S.Marketing to Maximize Profitability[J].The Journal of Business Strategy,1992,13(6):10-13
[2] JIAWEI H,MICHELINE K.Data Mining:Concepts and Techniques[M].San Francisco:Morgan Kaufmann,2000
[3] 叶孝明,黄祖庆.基于数据挖掘的零售业客户细分研究[J].现代管理科学,2006(6):63-64
[4] 周颖,吕巍,井淼.基于数据挖掘技术的移动通信行业客户细分[J].上海交通大学学报,2007(7):1142-1145
[5] CHING H C,YOU S C.Classifying the Segmentation of Customer Value via RFM Model and RS Theory[J].Expert Systems with Application,2009(36):4176-4184
[6] 徐翔斌,王佳强.基于改进RFM模型的电子商务客户细分[J].计算机应用,2012,32(5):1439-1442
[7] HUGHES A.Strategic Database Marketing:the Master Plan for Starling and Managing a Profitable Customer-based Marketing Program[M].New York:McGraw-Hill Professional,1994
[8] CHANG H C,TSAI H P.Group RFM Analysis as a Novel Framework to Discover Better Customer Consumption Behavior[J].Expert System with Application,2011,38(12):57-63
[9] 刘英姿,吴昊.客户细分方法研究综述[J].管理工程学报,2006,20(1):53-56
[10] 陆添超,康凯.熵值法和层次分析法在权重确定中的应用[J].电脑编程技巧与维护,2009(22):19-20
[11] 谢邦昌.数据挖掘基础与应用[M].北京:机械工程出版社,2012
[12] HSU F M,LU L P,LIN C M.Segmenting Customers by Transaction Data with Concept Hierarchy[J].Expert System with Application,2012,39(6):6221-6228
[13] MAHBOUBEH K,KIYANA Z.Estimating Customer Lifetime Value Based on RFM Analysis of Customer Purchase Behavior:Case Study[J].Procedia Computer Science,2011(3):57-63
[14] 曾小青.基于消费数据挖掘的多指标客户细分新方法[J].计算机应用研究,2013,30(10):2944-2947

Research on Customer Segmentation Method in Retail Industry Based on Data Mining

CAI Jiu-lin, ZHANG Lei, ZHANG Qiu-lan

(School of Management Science and Engineering, Qingdao University, Qingdao 266001, China)

Abstract: Due to the problem in the roughness of customer segmentation indicator and low accuracy in retail industry, a customer segmentation method in retail industry is propoesed on the basis of clustering analysis of data mining, and a set of RFM based on multi-indicator customer segmentation index system is constructed by using entropy value method to give indicator weight and then by using *K*-Means algorithm to conduct customer segmentation. Empirical research results show that this method is better than the traditional RFM based on segmentation method in the perspective of distinguishing capacity for customer behaviors feature and clustering compactness, and this method, with feasibility and validity, can better solve the problem in customer segmentation in retail industry and improves the customer relation management and marketing decision-making quality.

Key words: customer segmentation; RFM; entropy value method; *K*-Means

责任编辑:田 静
校 对:李翠薇